Modeling the AIDS Epidemic

Peter J. Denning

22 Sep 1988

RIACS Technical Report 88.27

NASA Cooperative Agreement Number NCC 2-387

Research Institute for Advanced Computer Science

Modeling the AIDS Epidemic

Peter J. Denning

Research Institute for Advanced Computer Science
NASA Ames Research Center

RIACS Technical Report 88.27
22 Sep 1988

The AIDS epidemic expands relentlessly. In October 1987, the President of the United States requested a national integrated scientific modeling effort to evaluate data already available and guide further data collection to reduce the uncertainties in estimates of prevalance and rates of spread of HIV. In July 1988, a workshop was jointly sponsored by the Office of Science and Technology Policy, Department of Health and Human Services, Department of Energy, and the National Science Foundation in response to this directive. The workshop recommended a variety of measures that add mathematical modeling to the arsenal of weapons we are developing to defeat HIV, foster and promote collaboration between modelers and other scientists, and encourage individuals and institutions to share data. Perhaps the most important result of the workshop was a transformation in the way the nearly 100 participants look at the AIDS question: they left with a much broader, community-oriented perspective.

This is a preprint of the column *The Science of Computing* for
American Scientist 76, No. 6 (November-December 1988).

Work reported herein was supported in part by Cooperative Agreement NCC 2-387
between the National Aeronautics and Space Administration (NASA)
and the Universities Space Research Association (USRA).

Modeling the AIDS Epidemic

Peter J. Denning

Research Institute for Advanced Computer Science

22 Sep 1988

In an earlier column (*1*), I discussed the use of computers to evaluate sophisticated mathematical models of the spread of AIDS. The epidemic shows no signs of abating (*2*). Since 1981, 68,000 people in the US have contracted the disease and as many as 1.5 million others are estimated to be asymptomatic carriers of HIV.† Efforts along the lines I indicated before are more important than ever.

In the early 1980s Robert May and Roy Anderson at Princeton and Klaus Dietz at the University of Tubingen began to develop mathematical models that would predict the spread of the disease. Around the same time, Stirling Colgate of the Los Alamos National Laboratory, who had come to believe that a disease that subverts the immune system is a more serious threat to humanity than nuclear war, began to promote internally Los Alamos's involvement in mathematical modeling of the disease; this resulted in the

†In the United States, the disease is transmitted primarily by sexual contact, and secondarily by sharing of blood such as with intravenous drug abuse and mother-to-child prenatal contact. There is no cure. Although vaccines are now under test, they are unlikely to protect more than a fraction of the population, and no generally protective agent is known. Recent studies have shown that the standard HIV seroconversion test may miss a fraction of those infected, because the virus may hide out in human cells for up to 18 months or more before inducing the formation of the antibodies detected by the test. The course of the epidemic is exceedingly difficult to project because of the extremely long asymptomatic incubation period of the disease, now estimated to average nearly 10 years.

organization in 1986 of a small group of applied mathematicians led by Mac Hyman. All these researchers share the goal of greater understanding of the dynamics of the disease through computer programs that are capable of forecasting the number of people with HIV under various assumptions. Not only would such programs help determine the extent of the disease if it were allowed to proceed unchecked, but more important, they would help evaluate the probable consequences of various intervention strategies, such as sex education and clean-needle programs for intravenous drug users.

Although the modeling effort begun by these few people represents a tiny fraction of the nearly $1.3 billion to be spent on AIDS by the US government in 1989, the importance of modeling was recognized recently through a workshop called the National Scientific Effort on AIDS Modeling and Epidemiology, held last July in Leesburg, Virginia. The workshop was convoked in response to presidential directives that charged the Office of Science and Technology Policy and the Department of Health and Human Services to reduce the uncertainties in estimates of prevalence and rates of spread of HIV. The Department of Energy and the National Science Foundation joined as cosponsors. The organizers, including Hirsch Cohen of IBM and Scott Layne of Los Alamos, brought together one hundred leading experts from several countries and from a diverse set of disciplines, who considered how the knowledge and resources at their command could be brought to bear on questions implied by President Reagan's directives: How accurate are the data we have now on the prevalence of the disease and the HIV infection? How can the existing research data be made available for use in computer models of the disease? How sharp is our understanding of the biological mechanisms of infection and the social processes that help transmit it? How does changing public awareness of the disease affect transmission? How can computer-based models be made available

for general use within the AIDS research and public health communities?

The workshop was divided into six groups, which met individually and in pairs. In what follows I will summarize their findings and recommendations.

The modeling group, chaired by James Glimm of the Courant Institute of Mathematics, focused on existing AIDS modeling efforts, areas of uncertainty, and data needs. The members of the group said that the goals of modeling in the next three years should be improved statistical methods and differential equations for estimating the prevalence of the HIV among the population, evaluation of proposed prevention strategies, short- and long-term projections of AIDS prevalence for health planning, and software tools for epidemiological studies. Statistical models, which extrapolate trends from past data, appear to be valid for predictions of up to five years. Differential equation models take into account contact and transmission among different groups of the population and are capable of predictions over much longer periods; their parameters include group sizes, HIV prevalence in each group, rates of partnership formation and dissolution, and infectivity as a function of time after contraction of the virus. The data needs are much more imposing for the differential equation models than for statistical models; a sizable new effort should be devoted to defining and obtaining appropriate data, making data from various studies available for use in models, and encouraging collaboration between modelers and those who gather data.

The biology and epidemiology group, chaired by James Curran of the Centers for Disease Control, discussed the key biological questions that must be answered to model the infectivity and transmission rates of the disease. Like studies of other sexually transmitted diseases, scientific studies of AIDS proceed in the midst of a charged social,

ethical, and political atmosphere. An amazing amount of knowledge about the disease has been gathered since its discovery in 1981. According to members of this group, models explain its generic complexities support research in a variety of ways, including the design of surveys, experiments, and samples, the study of small core groups that tend to drive the disease, the evaluation of medical and other intervention strategies, and the analysis of geographic differences.

The behavior and demography group, led by David Kanouse of the Rand Corporation, focused on the characteristics of different identifable groups such as homosexuals, heterosexuals, and intravenous drug users and on formations of sexual partnerships. The members examined how to translate the need for model parameters into studies that would provide data about five matters of concern: the core groups driving the disease; contact rates by region, age, sexual preference, and other factors; changing patterns over time, not just cross-sectional snapshots of the population; the relative size of each group in the general population; and the effective granularity of analysis -- individuals, partnerships, groups, or the whole population. They identified a large number of variables pertaining to sexual behavior and intravenous drug use, as well as gaps in information about high-risk populations.

The data collection and population sampling group, chaired by Lincoln Moses of Stanford University, concerned itself with the difficulties inherent in obtaining accurate data from people who may want to avoid identification. They considered sampling methods that might be used to overcome the reluctance to discuss one's behavior. Members of the group examined two kinds of data sources, the large national surveys and specialized studies. Three national surveys will produce data about HIV prevalence: the

national household HIV seroprevalence survey, planned for 1989, will request voluntary anonymous blood samples from 50,000 people; the national health and nutrition evaluation survey, also planned for 1989, will reach 30,000 persons; and a survey of women giving birth in 1988 will cover 30 states. The first two of these studies may be of marginal value because of the bias introduced by nonresponse. In one recent pilot study the same group was sampled in two different ways, once by a blind 100% sample and again by a direct interview; over half the HIV positive people were among the 18% nonrespondents. Thus modelers and epidemiologists need to stay within the realm of answerable questions and verifiable answers.

The data management and access group, led by Gio Wiederhold of Stanford University, considered standards for datasets that are made available for use by other researchers, methods to insure confidentiality for the individuals represented in the data, and means to make the data more conveniently available. This group's members recommended that a comprehensive directory of databases be assembled (*3*) and that owners of useful databases be offered financial incentives to make the information available. Large data collections, such as those assembled by the Centers for Disease Conttrol and the national surveys described above, should be expanded and improved, and those who maintain them should exploit the latest technological changes, such as high speed networking, distributed access to databases, software distribution, and new methods of publication and distribution such as compact optical disks and hypertext. Finally, the group recommended that guidelines be developed to insure privacy of all data.

The implementation and management group, chaired by Truman Brown of the Fox Chase Cancer Center, focused on the practical issues in implementing the scientific pro-

gram recommended by the other five groups in the context of the governmental and independent organizations conducting the research. Members examined ways that a coordinated program in modeling and data analysis could be implemented and managed within the government, considering a variety of options that would integrate modeling efforts with other AIDS research, foster interagency cooperation, and establish an oversight mechanism. They also considered incentives for collaboration among modelers, epidemiologists, biologists, and social, behavioral, and data management scientists.

The draft report was delivered about two weeks later to the Director of OSTP. This speedy conclusion was made possible by the presence of two dozen personal computers, allowing the six groups to complete their drafts before leaving.

The most significant result of the meeting was the shifts in understanding and the growth of mutual respect as the week progressed. Early in the week, many participants expressed suspicion by asking pointedly, "It looks like you modelers are telling us to provide you with our data. What will you do for us?" In a joint session at which the modelers described infectivity functions, the biologists suggested in the ensuing discussion a possible modification in light of recent results in virology suggesting that peaks of infectivity may occur in correlation with other infections in the host. At the end of the session, one formerly skeptical biologist declared with pleasure that he now had a much greater appreciation for the epidemic model and looked forward to further collaboration between biologists and mathematicians. By the end of the week, the modelers were accepted in the ranks of researchers arrayed against AIDS. Many had already agreed to collaborations with epidemiologists, biologists, and social and behavioral scientists, expressing a desire to participate in the field studies in which data are gathered.

Not only were mathematicians accepted by other AIDS researchers, but in general the separateness of the several communities represented by the working groups was overcome. Early in the week, the groups functioned as autonomous bodies, but by the end of the week they spoke of the larger community to which they belonged, declaring their intentions of working together to defeat the common enemy, AIDS.

A third shift concerned the willingness of individual scientists and institutions to share data. At the beginning of the workshop, participants were referring defensively to the scientific tradition that the investigator is responsible for the quality and integrity of results and also receives credit for them. This tradition has been manifested as a strong interest in intellectual property rights. Why should a scientist release his data before he himself has fully analyzed them and published the results? Why should he give away the lifeblood of his scientific career? When some suggested that there need to be ways to give credit for publishing a dataset in an on-line database or on optical disk, others responded that tenure committees would never recognize such action as legitimate publication. Someone expressed frustration over the dilemma between the tradition of science and the special needs generated by the AIDS crisis, exclaiming, "What a time for people's egos to get in the way!" Many participants felt the dilemma personally but by the end of the week they had discovered a variety of practical means of sharing data, including collaborative projects, a directory of available datasets, and various forms of electronic distribution.

I have little doubt that these shifts were made possible by the pairwise group sessions, a brilliant stroke by the workshop's organizers. These sessions enabled differences between the groups to be brought into the open, discussed, and resolved.

I would like to conclude with my own view of how computing technology can provide a new context within which many of the data management problems that motivated this meeting will be solved. I say this recognizing that the single greatest barrier to more effective AIDS research is not technological -- it is the unwillingness of people and institutions to share information. Several new technologies that will be available to the scientific community in the next three to five years -- networking, CD-ROM (compact disk read-only memory), software distribution, and graphics -- should facilitate exchanges between those who model and those whose research produces data.

The National Research Internet, based on NSFNET and a system of regional networks, will be operational within three years, providing network services to virtually all scientific communities. Databases can be attached to this network as resources to be used by any authorized scientist at any location. Servers associated with the databases can can distribute copies of portions of them and provide answers to queries.

With scientific journals stored on CD-ROMs that are readable on any personal computer with an optical disk scanner, the full data on which the findings of a paper are based can be included as an appendix at only a small increment of cost. The new hypertext systems allow the paper and the data to be stored in linked segments that can be brought into view simply by clicking on the text at appropriate places. For example, clicking on a portion of a table might bring up a window containing all the numbers used to derive the values in it, while clicking on an equation in the text might bring up a window containing its derivation. The programs developed for modeling the spread of disease can be made available for distribution on floppy disk or by network file transfer. Any researcher can then check the models against his own data.

The importance of innovative graphics derives from the fact that much of the existing data about HIV infection is spatial. The eye can readily appreciate a snapshot of the current distribution of the disease displayed on a map of the United States in various colors. Animated sequences of maps will reveal trends visually. Results of model calculations can be similarly displayed, and it is likely that many other aspects of the data can be understood more readily by visual means.

These are all new tools that will soon be available. For the present, the Leesburg workshop brought about a profound change in the way the participants looked at the AIDS question: they left with a much broader, more community-oriented perspective. Collaboration opened up among scientific communities and government agencies that have not regularly worked together. The capabilities of mathematical modeling have been added to the arsenal of weapons we are developing to defeat HIV. Things shifted in Leesburg.

References

1. P. J. Denning. 1987. "Computer models of AIDS epidemiology." *American Scientist* **75** (July-August). 347-351.

2. J. W. Curran, H. W. Jaffee, A. M. Hardy, W. M. Morgan, R. M. Selik, T. Dondero. 1988. "Epidemiology of HIV infection and AIDS in the United States." *Science* **103** No. 3 (February 5). 610-616.

3. S. Layne, T. Marr, S. Colgate, M. Hyman, and A. Stanley. 1988. "The need for national HIV databases." *Nature* 333 (June 9). 511-512.

4. R. M. May and R. M. Anderson. 1988. "Epidemiological parameters of HIV infection." *Nature* 333 (June 9). 514-519.

5. R. M. May and R. M. Anderson. 1987. "Transmission dynamics of HIV infection." *Nature* 326 (March 12). 137-142.

Box 1 -- Models of HIV Spread

The goal of a mathematical model is to predict the future state of a system or some future aggregate property of the system. One type of model, based on extrapolation of statistical trends, is useful for projections up to five years. An example is the relationship between HIV infection, incubation period, and AIDS (see ref. *1*). The other type of model is a system of differential equations that relate variables associated with a set of possible system states at each time t. A numerical solution of the equations yields the behavior of each variable over time.

A state in the second type of epidemiological model is defined by the combination of specific values for age, sexual preference, drug use, number of recent partners and their ages, stage of the disease in each partner, presence of other infections, and geographical location. The total number of states for a full geographical model of the United States might be on the order of 1 billion. Let s denote any one of these states.

Suppose that $n(s,t)$ denotes the number of people in state s at time t. Consider a set of equations of the form

$$n(s,t+dt) = n(s,t) + \sum_{s'} n(s',t)r(s',s)dt - \sum_{s''} n(s,t)r(s,s'')dt$$

where $r(s,s')$ denotes the rate of transition from one state s to another state s'. These equations would yield an exact solution for all $n(s,t)$ at all t, given the initial conditions at time 0 and the transition rates.

In practice, however, these equations cannot be solved directly. Where do the transition rates $r(s,s')$ come from? In a realistic model, there may be 10^9 states, giving rise to 10^{18} rates. No conceivable study of the disease can possibly be expected to take that many measurements.

Thus the modeler is forced to map a relatively small number of available measurements onto the large set of unknown rate coefficients and thereby obtain an approximate solution. The calculations from the resulting equations can be extremely sensitive to how the limited information gets distributed among the coefficients. An example of an assumption is that the transmission rate between an arbitrary pair of states depends only on the infectivity function associated with the source state of the pair. Great care is needed in formulating these assumptions; the design of the model must proceed in concert with the design

of experiments to gather data.

The problem becomes even more complicated when there is feedback within the system such that some of the coefficients depend on how many people are in certain states. For example, the rate at which one person gets infected by a randomly chosen partner will depend on the fraction of the population already infected. Now the results of the model depend also on the assumption used to specify the coefficients affected by feedback. The uncertainty in the results requires extensive validation of models before they can provide useful information to AIDS disease-control policymakers.

Box 2 -- The Reproductive Factor

A disease will continue in a population if each infected individual infects at least one other before he leaves the population. The reproductive factor R measures this (*4*). In epidemiology it is defined as

$$R = CTD$$

where C is the contact rate, T is the probability of transmission per contact, and D is the duration of infectivity. In the case of AIDS, for example, an HIV-infected person with 200 sexual contacts per year, a 1% probability of transmitting the infection per contact, and a 10-year incubation period would have $R = 20$. Thus the infected person leaves the population with 20 others infected, a net gain of 19. The goal of intervention strategies is to reduce R to less than 1, in which case the disease will eventually die out. Computer-based models can be used to discover whether a proposed strategy is likely to accomplish this. Thus far, models have confirmed what is already obvious: people must significantly change their behavior to quell the epidemic.

Box 3 -- Infectivity

May and Anderson have suggested a simple model of infectivity versus time (*4,5*). Let $I(t)$ denote the infection rate at time t since infection with HIV. An initial peak of $I(t)$ may last a few months and be on the order of 10% probability of infection per sexual contact, followed by a low trough of perhaps 0.1% for seven or eight years, and then a gradual rise to 10% and higher as the AIDS symptoms begin to appear.

The Los Alamos model has shown that the onset of the catastrophe caused by the epidemic (defined as a signficant portion of the population having HIV) is sensitive to the assumption for $I(t)$. If $I(t)$ is a constant with the same mean as the May-Anderson function, the catastrophe might take 25 years rather than 20 years to arrive. If the initial peak of infectivity is given more prominence, the catastrophe might occur in 15 years. It is thus apparent that an understanding of the biology of infection is critical to any model.

Recent work in virology reported by Scott Layne, John Spouge, Micah Dembo, and Peter Nara sheds new light on the model for $I(t)$. The probability of infection by HIV is directly related to the concentration of T4+ (white blood) cells at the site where the virus is introduced. If the host has any infection that increases the number of cells in the urinary tract -- for example, prostatitis or clymidia -- he or she may have a signficant increase in susceptibility. Thus the presence of other diseases -- cofactors -- in the host may generate other peaks of infectivity not in the simple model.

Box 4 -- The workshop

The National Scientific Effort on AIDS Modeling and Epidemiology was organized by the Office of Science and Technology Policy, the Department of Health and Human Services, the Department of Energy, and the National Science Foundation in response to directions from the President in October 1987 that sought an integrated scientific modeling effort to evaluate data already obtained and guide further data collection to reduce the uncertainties in estimates of prevalence and rates of spread of HIV (the AIDS virus). The specific goals of the workshop were to:

1. Review existing modeling efforts and identify classes of models that are appropriate for a national program.
2. Identify the basic information in all domains -- biological, behaviorial, demographic -- required to parameterize the models.
3. Review status of the existing knowledge base about HIV in the US.
4. Identify areas in which uncertainty in the data or the models makes a significant difference in understand the epidemic; review data collection efforts and approaches to reducing this uncertainty.
5. Estimate resource requirements for data collection and management.
6. Identify critical administrative and management issues involved in implementing a program and develop approaches to resolve them.

Nearly 100 scientists from a diverse set of discplines and several countries participated in the meeting. They were divided into six groups that met individually and in pairs.

The group findings were distilled into the major findings of the workshop: Mathematical and statistical modeling must be brought fully to bear on the AIDS epidemic, as added technologies to the field. Models of other sexually transmitted diseases are only partially applicable. AIDS incidence data allows only short range projections. Progress toward long range projections requires new biological and behavioral data and a collaborative effort among modelers, epidemiologists, biologists, and social and behavioral scientists. A pool of very high quality mathematical talent exists to conduct the modeling effort.

Access to existing and future data must be improved.

Similarly the group recommendations were distilled into the recommendations of the workshop: In the area of funding, support for AIDS modeling should be increased. Priority should be given to projects that include collaboration between modelers, epidemiologists, biologists, and social and behavioral scientists, and to projects that will make their data available for others to use. Funding agencies should include modelers on their staffs and establish targeted programs for modeling. In the area of data access, a comprehensive directory of available AIDS datasets should be established and maintained and placed online. Access to the large public datasets should be expanded, and small population datasets should be made available to qualified researchers. The national surveys should go forward. The "new friendships" between modelers, epidemiologists, biologists, and social and behavioral scientists that emerged at the workshop must be nurtured through more interagency and interdisciplinary meetings and through the professional societies.